Exploratory Data Analysis to Detect Preterm Risk Factors

Jonathan C. Prather, M.S.[1], David F. Lobach, M.D.,Ph.D.,M.S.[1], Joseph W. Hales, Ph.D.[1], Linda K. Goodwin, R.N., Ph.D. [1], Marvin L. Hage, M.D. [1], David A. Nagey, M.D., Ph.D. [2], and W. Ed Hammond, Ph.D.[1]

[1]Duke University Medical Center, Durham, North Carolina
[2]The Johns Hopkins University School of Medicine, Baltimore, Maryland

This poster is a progress report of a project to mine a computerized patient record for new medical knowledge in perinatal health care. Preterm birth is responsible for a majority of the morbidity and mortality associated with the perinatal period of infants in the United States. Early identification of women at risk for preterm birth allows for pharmacologic and behavioral interventions that attempt to prolong gestation. Unfortunately, current risk assessment scoring tools are only 17 to 38% accurate in correctly predicting preterm birth.[1] The inadequacy of current risk scoring tools and preterm birth prevention programs likely stems from a failure to fully identify factors that cause preterm birth. The hypothesis of our project is that novel approaches to data analysis applied to an extensive clinical database will detect previously unrecognized factors that contribute to preterm birth. Identification of these factors will significantly improve the ability to identify mothers at risk for preterm delivery. Early and accurate identification of these mothers will allow interventions to be targeted at women with the greatest risk for preterm delivery.

In this study, clinical data in the Duke University Medical Center Perinatal Databank are analyzed to detect factors associated with preterm birth. This databank contains data collected in a regional perinatal computerized patient record and represents one of the largest and most comprehensive sources of perinatal data in the United States. It consists of detailed, coded clinical information on the prenatal, intrapartum, and postpartum care associated with over 22,000 births from a six county region. Coded and numerical clinical data are extracted from the databank,[2] formatted and cleaned for data analysis, and categorized into temporally stratified data sets. These data sets contain gestational ages at delivery and clinical observations generated during the preconception, first trimester, second trimester, and third trimester stages of pregnancy. After applying data reduction techniques of principal component analysis and exploratory factor analysis, component and latent factor scores are calculated for each birth. The raw data sets, sets of component scores, and sets of latent factor scores are then used to individually create prediction models using multiple linear regression, Cox regression, and neural networks. The resulting models are tested against both one year of Duke data not used in developing the models and a similar perinatal data source originating at the University of Maryland Medical Center. The sensitivity and specificity of each prediction approach will be compared using receiver operating characteristic (ROC) analysis, and the modifiable risk factors will be identified.

Newly discovered relationships found in computerized patient record systems will potentially lead to better understanding between observations and outcomes in perinatal care and other fields of medicine. After more factors contributing to preterm birth are identified by mining clinical data from the prenatal course of preterm infants, revised risk prevention programs can be created. Perinatal health care providers will potentially have more accurate predictive models to assess preterm risk. Computerized patient record systems (CPRS) now contain enough detailed clinical data to help us find the relationships between multiple variables and the outcomes they influence without the traditional costly and labor intensive data collection for research.

This work supported in part by AHCPR dissertation grant R03 HS09331-01.

1. McLean, M., Walters, W., and Smith, R. (1993) Prediction and early diagnosis of preterm labor: a critical review. Obstetrical and Gynecological Survey. 48 (4):209-225.
2. Prather JC, Lobach DF, Hales JW, Hage ML, Fehrs SJ, Hammond WE. (1995) Converting a Legacy System Database into Relational Format to Enhance Query Efficiency. Proceedings Annual Symposium Computer Applications Medical Care, 19:372-376.

1091—8280/97/$5.00 © 1997 AMIA, Inc.